








Research Article

# Influence of storage time on the stability of diatom assemblages using DNA from riverine biofilm samples

Jonathan Warren<sup>1</sup>, Sean Butler<sup>1</sup>, Nick Evens<sup>1</sup>, Laura Hunt<sup>1</sup>, Martyn Kelly<sup>2,3</sup>, Lindsay Newbold<sup>4</sup>, Daniel S. Read<sup>4</sup>, Joe D. Taylor<sup>4</sup>, Kerry Walsh<sup>1</sup>

<sup>1</sup> Environment Agency, Horizon House, Deanery Rd, Bristol, BS1 5AH, UK

<sup>2</sup> Bowburn Consultancy, 11 Montaigne Drive, Bowburn, Durham, DH6 5QB, UK

<sup>3</sup> School of Geography, Nottingham University, Nottingham, NG7 2RD, UK

<sup>4</sup> UK Centre for Ecology & Hydrology, Wallingford, OX10 8BB, UK

Corresponding author: Jonathan Warren ([jonathan.warren@environment-agency.gov.uk](mailto:jonathan.warren@environment-agency.gov.uk))

This article is part of:

***Towards standardized molecular biodiversity monitoring***

*Edited by Teresita Porter, John Darling, Kelly Goodwin, Tiina Laamanen, Kristian Meissner, Toshifumi Minamoto*



Academic editor: Tiina Laamanen

Received: 1 July 2024

Accepted: 7 August 2024

Published: 22 August 2024

**Citation:** Warren J, Butler S, Evens N, Hunt L, Kelly M, Newbold L, Read DS, Taylor JD, Walsh K (2024) Influence of storage time on the stability of diatom assemblages using DNA from riverine biofilm samples. *Metabarcoding and Metagenomics* 8: e129227. <https://doi.org/10.3897/mbmg.8.129227>

Copyright: © Jonathan Warren et al.

This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

## Abstract

DNA sequencing of diatom assemblages from biofilms has already been used to assess the ecological status of freshwater in the UK. However, recent work using DNA data from these biofilms suggests that alternate metrics that capture the broader taxonomic and functional information to demonstrate importance of microbial biofilms could be useful. Exploring this potential requires large numbers of samples over time and space to be analysed. Sample archives could be used to meet this need, but the compositional stability of microbial communities in stored biofilm samples for more than one year is uncertain.

This study compared changes in diatom assemblage structure using metabarcoding analysis of river biofilm samples before and after storage at -20 °C in an RNAlater-based nucleic acid preservative. We found minimal changes in the diatom assemblages in the samples when stored for up to three years. Slight differences in certain groups observed resulted in four samples changing ecological status. However, the overall differences were not significant across replicates, suggesting any genuine differences in assemblages are likely masked by sub-sampling, PCR, or primer biases. These findings are similar to those observed in other studies looking at variations between analysts and sequencing instruments. This indicates that the diatom assemblages in the archived biofilm samples are stable. This will give greater confidence that archived samples can be used for further research, including exploring broader microbial taxa and their responses to environmental change, potentially leading to the development of reliable microbial metrics for integration into biomonitoring programs.

**Key words:** Biomonitoring, diatom assemblage, metabarcoding, microbiome sample preservation, surveillance

## Introduction

In the United Kingdom, environmental regulators, including the Environment Agency of England, employ various methods using biological indicators to assess the ecological status of lakes, rivers, and estuaries. One method involves the use of diatoms from biofilms in rivers and lakes as proxies for wider

phytobenthos (Kelly et al. 2008). This assessment utilises Ecological Quality Ratio (EQR) metrics (European Parliament 2000) based on the Trophic Diatom Index (TDI) to assess ecological status. Classifications are determined based on the evaluation of the diatom assemblage composition using either light microscopy or high throughput sequencing (Kelly et al. 2008; Kelly et al. 2020).

In addition to the diatom assemblage, biofilms support a diverse microbial community embedded within a slimy matrix, which promotes their growth and survival (Neu and Lawrence 1997). Biofilm communities include bacteria, fungi, a diversity of algae, and protozoa that all contribute to important ecosystem processes, such as primary productivity, decomposition, biogeochemical cycling, and pollutant degradation (Falkowski et al. 2008; McGuire and Treseder 2010; Mishra et al. 2021). In many European countries, microbial bioindicators have largely been restricted to diatoms, phytoplankton, and targeted bacterial groups, such as faecal indicator bacteria, for routine ecological assessments under the Water Framework Directive (European Parliament 2000) and the Bathing Water Directive (European Parliament 2006). Whilst there is recognition that a much more diverse microbial community exists within river biofilms, we are not fully exploiting this information to build new indices and metrics (Environment Agency 2023; Kelly et al. 2024).

The integration of microbial diversity and functional indicators into biomonitoring for the assessment of anthropogenic pressures has been advocated for many years, and has recently gained momentum (Jackson et al. 2016; Cordier et al. 2019; Sagova-Mareckova et al. 2021; Warnasuriya et al. 2023). There is a good foundation for the exploration of new diagnostic microbial metrics and indices (Sagova-Mareckova et al. 2021), enabled by advances in high-throughput sequencing and computational approaches, such as machine learning (Cordier et al. 2019; McElhinney et al. 2022) and network analyses (Codello et al. 2022; Guseva et al. 2022). Such approaches allow the interrogation of relationships between microbial communities, their attributes, and anthropogenic pressures, facilitating a greater understanding of microbiomes and their responses to environmental change (e.g., Deutschmann et al. 2021; Eastwood et al. 2023).

Currently, data and models are needed to identify reliable microbial bioindicators (Fontaine et al. 2023). Large spatial and temporally relevant microbial datasets are required to mine and identify candidate ‘features’ of microbial communities that have the potential to be developed and upscaled more widely as bioindicators of ecosystem function and predictors of change (Astudillo-García et al. 2019). Generating such datasets is costly due to the expense associated with extensive field sampling and analysis. Thousands of biofilm samples are collected across England’s river network as part of routine ecological assessments using diatoms. This presents a valuable archived resource to further explore wider microbial bioindicators, as samples are stored in a nucleic acid preservative and frozen at -20 °C (Kelly et al. 2020). A previous study by Baricevic et al. (2022) showed that preserved biofilm samples stored at -20 °C remained stable for 12 months. No significant impact was observed on DNA quality, yield, and the overall composition of phytobenthic diatom assemblages that reflected the site origin. However, the stability of archived biofilm samples over longer timescales remains unknown. With multiple years now in storage across thousands of sites river biofilm samples archived by the Environment Agency offer a unique opportunity to explore the spatiotemporal variability of the river microbiome across England in a cost-effective manner.

This study aimed to assess the stability of river biofilm samples for up to three years of freezer storage in preservative and used diatom assemblages as a proxy for the integrity of the overall microbial species in the biofilms. We assessed the suitability of reusing existing samples for the generation of large DNA datasets, which could enable the characterisation of the microbial response to environmental change.

## Methods

To assess the impact of storage time on diatom assemblages, samples collected and analysed in 2019 and 2020 for routine analysis that had been stored concentrated in preservative were re-analysed in December 2022. The reanalysis was performed twice so that differences due to storage could be disentangled from the expected stochastic differences due to sub-sampling.

### Sample collection and selection

River biofilm samples collected in 2019 ( $n=50$ ) and 2020 ( $n=14$ ) as part of the Environment Agency's routine monitoring of diatoms in rivers for ecological status assessments were used in this study. Samples were collected between April–November and analysed between September–January of the corresponding year. Samples were selected for reanalysis based on the total number of reads regardless of sequence taxonomic assignment passing quality control for routine analysis ( $>50,000$  reads), and to cover a broad spread of geographic regions across England. A total of 64 frozen biofilm samples were selected for re-analysis in late 2022, following storage in preservative for 2 or 3 years after the initial analysis (Fig. 1).

Sample collection was performed as previously described in Kelly et al. (2020). Briefly, biofilm-covered stones were collected in a tray and scrubbed with deionised or tap water using a clean toothbrush. Using a pipette, 5 ml of the biofilm suspension was transferred to a 15 mL tube containing RNAlater-based preservative (3.5 M ammonium sulphate, 17 mM sodium citrate, and 13 mM Ethylenediaminetetraacetic acid), transported to the laboratory via an overnight courier at  $5\pm 3$  °C, and stored frozen at  $-20\pm 5$  °C prior to DNA extraction.

### DNA extraction and diatom assemblage metabarcoding

All samples were processed, according to Kelly et al. (2020). Samples were thawed and then centrifuged at  $3000\times g$  for  $15\pm 2$  min at  $5\pm 2$  °C to form a concentrated biofilm pellet. The pellet was resuspended in 0.5 mL of the supernatant and the rest discarded. DNA from 0.1 mL of the resuspended pellet was extracted using the Qiagen blood tissue kit (#69506) with an extended overnight lysis step in a rocking incubator at  $56\pm 4$  °C. The target *rbcL* region was amplified by PCR using *rbcL*-646F (5'-ATGCGTTGGAGAGARCGTTTC-3') and *rbcL*-998R (5'-GATCACCTTCTAATTTACWACAACACTG-3') to generate an amplicon library for each sample. While these primers are used for the routine assessment of diatom assemblages, they also amplify other non-diatom phytobenthos (Kelly et al. 2024). The amplicon libraries were purified using solid-phase reversible immobilisation (SPRI) beads. After cleaning, the libraries were tagged using Illumina Nextera XT unique dual-indexed adapters (#20091654) and purified. Tagged and



**Figure 1.** Map showing the geographic spread of biofilm sampling points across England.

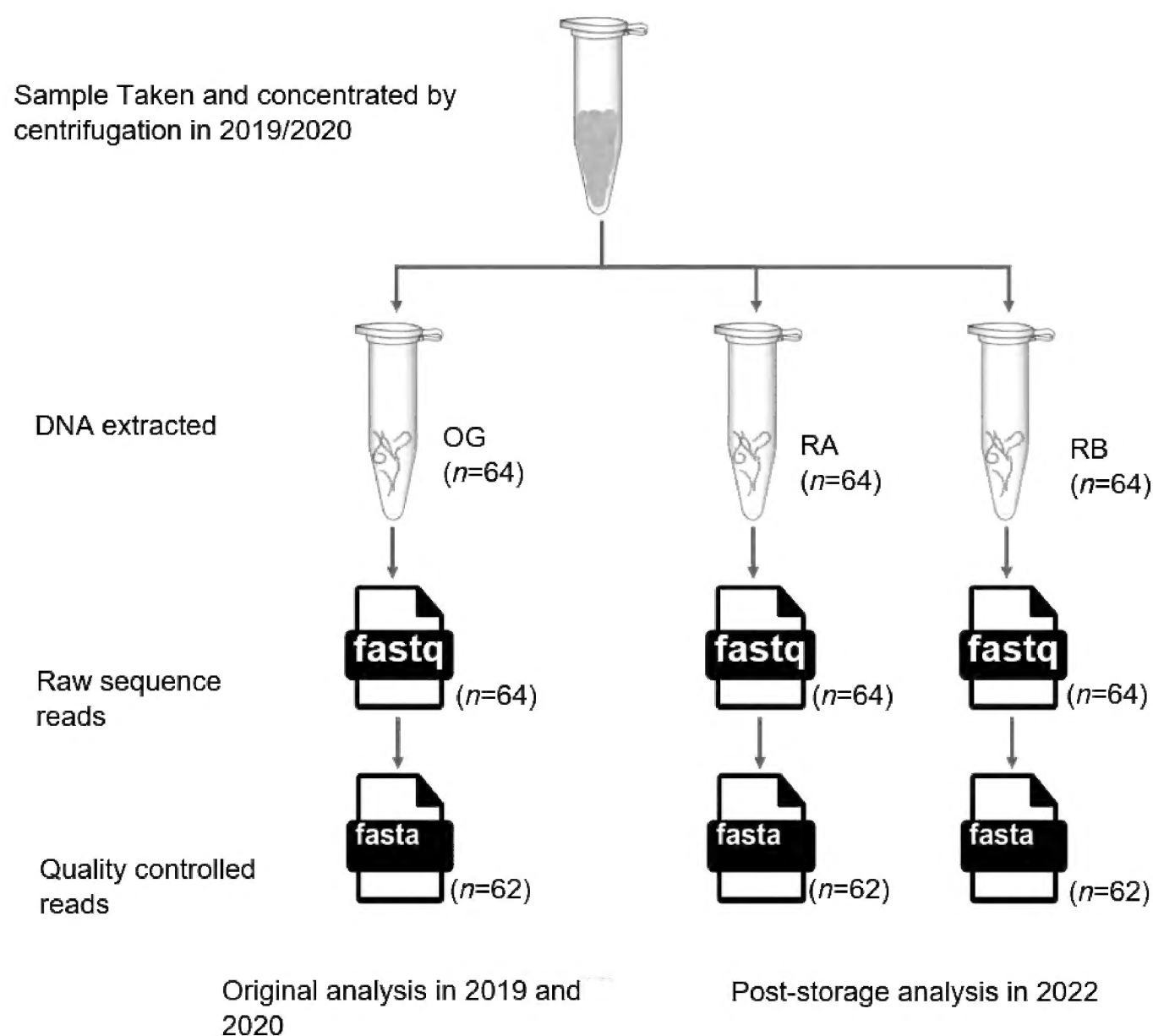
purified amplicon libraries for each sample were normalised and pooled using molecular grade water and sequenced on an Illumina MiSeq instrument using a 600 cycle V3 reagents kit (#MS-102-3003). This original analysis (OG) forms the first sub-sample group of this study. The complete PCR conditions and clean-up procedures are detailed in Suppl. material 1. The remaining concentrated biofilm sample was stored at  $-20 \pm 5$  °C in RNAlater-based preservative.

After storage, the samples were thawed for reanalysis by aliquoting two 0.1 mL sub-samples (RA and RB; Fig. 2) and processed following the same method described above. In total, 192 samples were sequenced across the three sub-sample groups.

### Bioinformatic analysis

Raw sequence reads were imported into the QIIME2 environment (v2022.8.3) (Bolyen et al. 2019). Primer sequences were removed using cutadapt, discarding reads with no matching primer sequences (Martin 2011). DADA2 was used to denoise, dereplicate, and remove chimeras on a per run basis table (Callahan et al. 2016). Data were combined into a single set of amplicon sequence variants (ASVs) and an abundance. Taxonomy was assigned to ASVs using QIIME2's scikit-learn multinomial naïve Bayes classifier against the diat.





**Figure 2.** Experimental design: samples were collected and processed in 2019 and 2020 (OG;  $n=50,14$ ). Post-storage samples were reanalysed in 2022 on two replicate sub-samples from each archived sample (RA and RB). Note that 2 samples originally taken in 2019 from group RB failed quality control and were removed from the study.

barcode reference database v11.1 (2022) with a 95% confidence threshold (Pedregosa et al. 2011; Bokulich et al. 2018; Rimet et al. 2019). After taxonomic assignment, data was not filtered so both diatom and other non-target phyto-benthos were included in all analyses.

### Statistical analysis

ASV abundance and taxonomy data were imported into R (v4.1.2) using qiime2R (0.99.6) and phyloseq (v1.38.0), and all data was visualised using ggplot2 (v3.4.2), ggvenn (v0.1.10), gghighlight (v0.4.0), and ggally (v2.1.2) packages (McMurdie and Holmes 2013; Wickham 2016; Bisanz 2018; Schloerke et al. 2021; R Core Team 2022; Yutani 2022; Yan 2023). Abundance data were not rarefied (McMurdie and Holmes 2014), but sequencing depth and diversity were investigated using rarefaction curves with the rarecurve function in vegan (v2.6-4) (Oksanen et al. 2022). For sub-samples in which the rarefaction curve did not plateau at the final read depth, all replicates of that sample were removed from the rest of the analysis. To compare trends in diversity across the dataset, distances were calculated at the ASV level using the Bray-Curtis distance measure. Bray-Curtis distances were used to account for variation in ASV detection and abundance between and within samples. Subsequently, distances were used to evaluate the influence of sample storage using permutational multivariate analysis of variance (PERMANOVA) with the adonis2 function from vegan (Oksanen et al. 2022).

The presence and absence of taxa were compared at both genus and species levels and visualised using ggvenn. The differential abundance of taxa was assessed using DESeq2 (v1.34.0), and relevant differences were visualised using ggplot and gghighlight (Love et al. 2014). The TDI, EQR, and classifications were calculated for each sample using darleq3 (v0.9.8) to assess the impact on ecological status classification (Kelly et al. 2020) and visualised using ggally.

The raw sequence files were deposited at the European Nucleotide Archive under the accession number PRJEB76460.

Results

Sample quality control

Two sub-samples failed quality control due to poor read depth, causing the rarefaction curves not to plateau, and as a result, all sub-samples for that sample were discarded. In total, 186 sub-samples from 62 samples passed quality control and were used for statistical analysis.

Across all samples, 12.9 million reads passed quality control. The number of reads per sample ranged from 4,980 to 258,229, with a median frequency of 64,151. A total of 5,138 unique ASVs were detected, and all were assumed to be phytobenthic algae; of these, 1,403 could be assigned to the species level, accounting for 59.3% of all reads, and a further 2,358 to the genus level, accounting for 80.8% of all reads in total, making these suitable levels with which to assess taxon detection (Table 1).

Table 1. Number of reads that the taxonomy was assigned to at each rank.

Taxonomic Level	Numbers of unique values at rank	Accumulative ASVs at level	Number of reads assigned to level or better	Percentage of total reads
ASVs	5,138	NA	12,905,482	100
Species	185	1,403	7,653,091	59.3
Genus	81	2,358	10,428,673	80.8
Family	38	2,491	10,809,236	83.8
Order	21	2,563	10,902,280	84.5
Class	8	2,793	11,202,988	86.8
Phylum	5	3,673	11,973,295	92.8
Kingdom	2	3,762	12,166,798	94.3

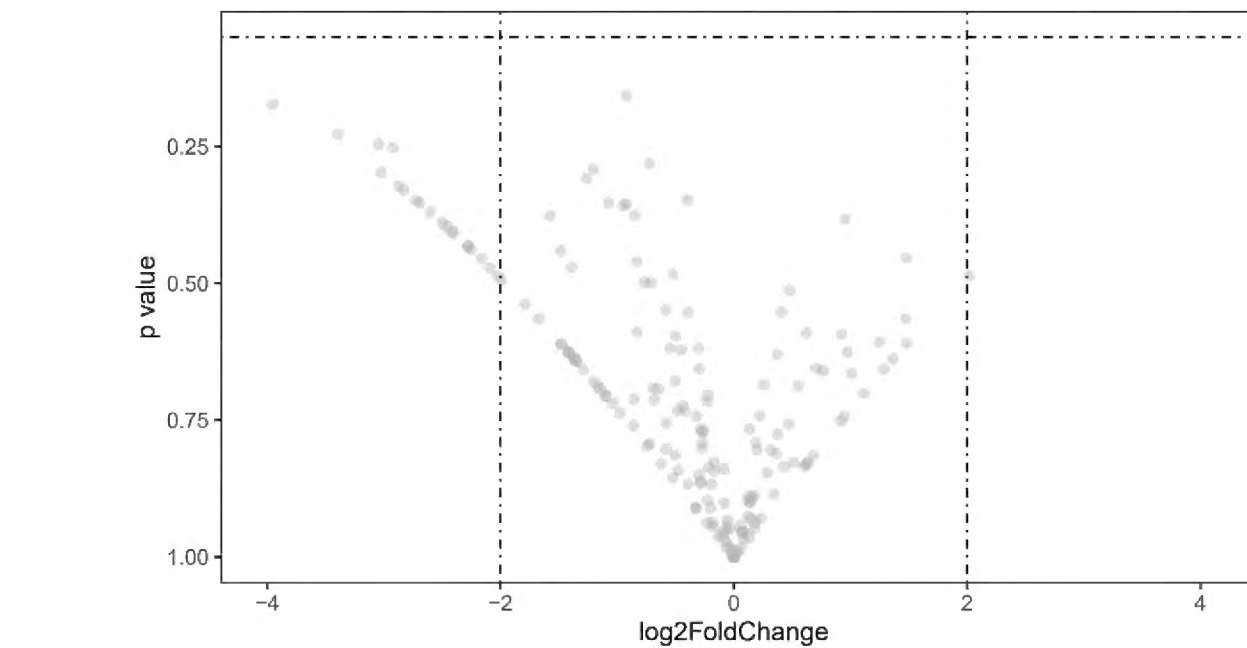
Impact of storage on diatom assemblages and taxon detection

At the ASV level, differences in diatom assemblages were mostly due to differences in the sample sites (PERMANOVA,  $R^2=0.383$ ,  $p=0.001$ ). Differences in storage conditions between the original and repeated sub-samples accounted for < 1% of the difference in diatom assemblages and were not statistically significant (PERMANOVA,  $R^2=0.005$ ,  $p=0.212$ , full model output in Suppl. material 2).

The abundance of taxa at the genus and species levels was not significantly different between the original and post-storage sub-samples. At a log2fold change no species or genera were detected at a significantly higher abundance post-storage (Fig. 3).

Most genera were detected in all replicate groups (84.0%; Fig. 4). However, seven genera were detected unevenly between the original and post-storage samples. All seven genera were detected only after storage and at low levels of detection (<1000 reads) across very few samples (1-3), whereas the original analysis had no unique genera (Table 2).

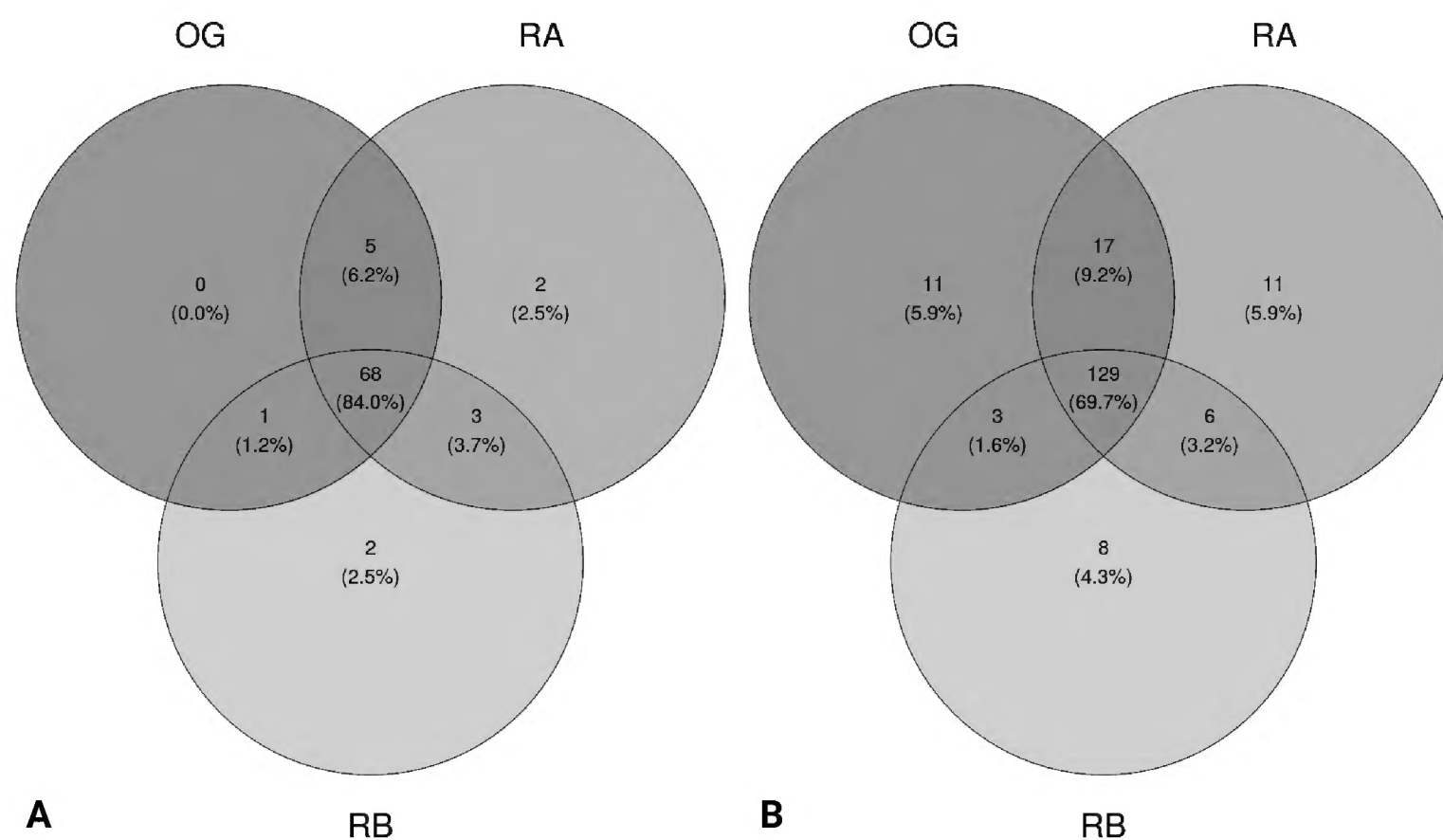
Similarly, at the species level 69.7% of species were detected in all replicate groups. However, 11 species were uniquely detected in the original samples and 24 were detected in the post-storage (RA and/or RB Fig. 4). Of the 35 species, all but three belonged to genera that were more widely detected (Suppl. material 4). *Surirella solea* and *Placoneis constans* were the only diatom species that did not belong to a more widely detected genus but were detected in the post-storage replicates only. In addition, the non-target species *Quercus robur* (common name: English Oak) was detected in three post-storage replicates only. For most of the species which were differentially detected, overall reads were low (<1000 reads) and were detected across very few samples (1-3). *Pinnularia viridis* was the only species detected at higher numbers (>1000), although this was possibly inflated by detection in a sample with an above-average read depth (181,687).



**Figure 3.** Volcano plot showing the fold change in abundance of the genus and species between the original and RA sample analyses. The horizontal dotted line denotes significance ( $p < 0.05$ ), and the vertical lines denote the magnitude of difference, where the lines are set at the equivalent of a 4-fold difference. Points to the right denote taxa found more abundantly in the original analysis, and points to the left denote taxa more abundant in the repeat analysis. Similar pairwise comparisons were made between all three groups (OG, RA, and RB), which also showed no significant difference (see Suppl. material 3).

**Table 2.** Uniquely detected genera between sample replicates. Numbers in parentheses represent the number of samples in which the genera were detected.

Genus (* indicates non-diatom)	Number of reads across the replicate group		
	OG	RA	RB
<i>Nupela</i>	0 (0)	131 (3)	106 (3)
<i>Bacillaria</i>	0 (0)	98 (1)	25 (1)
<i>Stenopterobia</i>	0 (0)	0 (0)	26 (1)
<i>Chaetoceros</i>	0 (0)	0 (0)	117 (1)
<i>Gedaniella</i>	0 (0)	9 (1)	0 (0)
<i>Placoneis</i>	0 (0)	3 (1)	0 (0)
<i>Quercus*</i>	0 (0)	34 (2)	29 (1)



**Figure 4.** Venn diagrams of genera (**A**) and species (**B**) common and unique to each sample type. OG is the original, and RA and RB are sub-samples analysed post-storage.

### Metric differences

TDI values generated from the diatom assemblages representative of the original and post-storage samples were generally similar. The mean difference in TDI before and after storage was 3.18, st. dev=7.33 (OG vs RB: 4.03, st. dev=9.22). The TDI values of all three replicate groups were highly correlated and statistically significant (Pearson's correlation coefficients ranged between 0.892 and 0.980;  $p < 0.001$  for all comparisons) (Fig. 5).

When assigning ecological status classifications using TDI scores, 75.8% of samples were assigned to the same class across all replicates (79.0% of samples were the same or one class different). When comparing any two of the replicate groups, the OG and RA groups had the highest agreement, with 90.4% of samples assigned to the same class, whereas RA and RB had the highest agreement at the same class or one class difference at 98.4% (see Table 3).

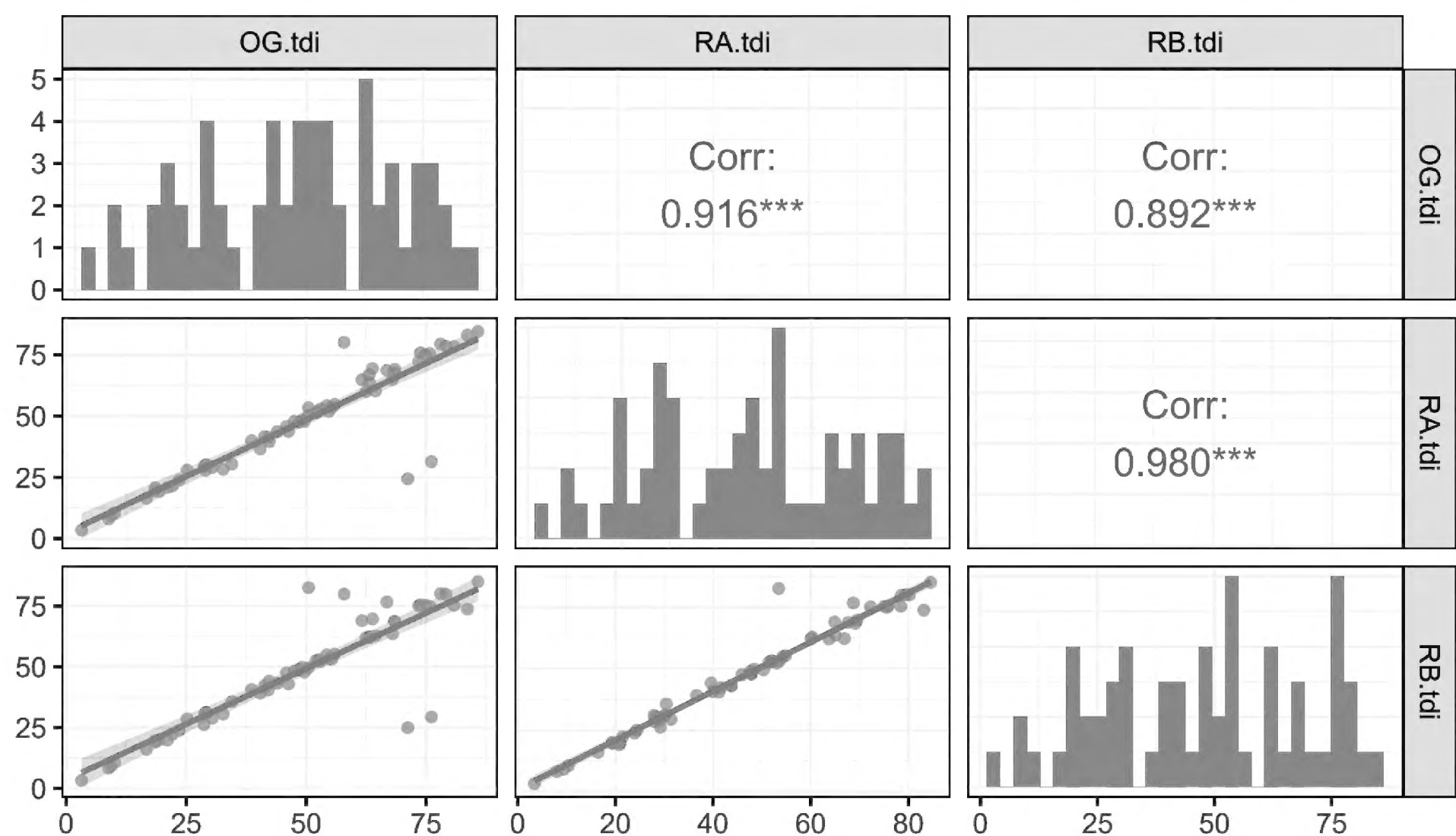
When comparing the TDI and the derived ecological status classifications, there were four clear outliers: S09, S43, S51, and S55. In one instance, RB (S09) was an outlier, and in the other three instances, the OG replicate was an outlier.

The relative abundance of diatoms between outlier sample replicates was compared using bar plots at the order level (Suppl. material 5). In sample 09, the RB replicate contained a much larger proportion of *Amphora pediculus* (47.1%; displayed as part of order Thalassiophysales in supplementary materials bar plot) than the OG and RA replicates (22.2% and 17.5%, respectively); otherwise, the communities appeared similar. However, in the other three outlier samples, the differences in community compositions were more apparent compared to non-outliers (See Suppl. materials 5, 6).

### Discussion

This study building on the work of Baricevic et al. (2022) improves our understanding of the impact of long-term storage on diatom assemblages. We examined samples frozen at -20 °C for up to three years in nucleic acid preservative





**Figure 5.** Correlogram of TDI scores showing differences between sample types and Pearson's correlation between scores. All Pearson correlations were highly significant ( $p < 0.001$ ). Histograms show the distribution of TDI scores across replicates.

**Table 3.** Matrix showing differences in the number of samples assigned to each ecological status class (bad, poor, moderate, good, high) between replicates. The samples assigned to the same class by two replicate tests are highlighted in green. Samples that were different by one class are highlighted in yellow. Samples in which the difference between replicates is greater than one class are highlighted in bold.

		OG					RA				
		B	P	M	G	H	B	P	M	G	H
RA	B	0	0	0	0	0					
	P	0	6	1	1	0					
	M	0	0	9	1	0					
	G	0	0	1	10	0					
	H	0	1	1	0	31					
RB	B	0	0	0	0	0	0	0	0	0	0
	P	0	4	4	2	0	0	6	3	1	0
	M	0	2	5	1	0	0	2	5	1	0
	G	0	0	2	8	0	0	0	2	8	0
	H	0	1	1	2	31	0	0	0	1	33

to assess the impact of storage on DNA recovery and to understand whether diatom assemblages exhibited any significant change in their composition. Diatom assemblages were used as a proxy for inferring impacts on the wider microbiome because they are the only microbial component of the biofilm currently assessed as part of the Environment Agency's routine monitoring program that has metabarcoding data to benchmark storage impacts. Archived samples collected as part of large routine monitoring programmes provide

unique opportunities to generate large comprehensive spatial and temporal datasets cost effectively to facilitate the development of new diagnostic metrics and indices.

The results of this study show that the storage of biofilm samples for up to 3 years in preservative frozen at -20 °C resulted in no significant difference in diatom assemblage diversity at the ASV level (PERMANOVA,  $R^2 = 0.005$ ,  $p = 0.212$ ), differences in 2 and 3 years of storage time were also compared and were equally insignificant (PERMANOVA,  $R^2 = 0.011$ ,  $p = 0.216$ ). Similar findings were found by Baricevic et al. (2022), where storage conditions of up to one year were insignificant. In both of these studies, it is likely that some of the small statistically insignificant differences were due to variability arising from analytical repeats; differences in diatom diversity at the ASV level observed in this study were similar to those reported by Kelly et al. (2018) when comparing differences observed between different analysts and different sequencing instruments.

Other studies have compared other microbial groups and storage conditions in similarly dense microbial sample types, reporting varying differences in the impact on the observed community. Tap et al. (2019) observed minimal effects of storage time on bacterial communities in faecal samples. Communities from samples stored for 5 years in preservative at -80 °C were comparable to the reference samples. Other studies have investigated the impact of storage conditions on bacterial communities in faecal material at shorter timescales, with similar findings (Bundgaard-Nielsen et al. 2018; Dully et al. 2021; Kim et al. 2023). However, Delavaux et al. (2020) found that the use of RNAlater significantly affected bacterial communities compared to samples frozen in liquid nitrogen and stored at -80 °C, but only at one of the two sites tested. They also investigated the impact on fungal communities but found no significant differences due to any of the storage conditions tested.

In the present study, when comparing individual taxonomic groups across replicate groups, there were no significant differences. At the species and genus levels, there were no statistically significant differences in the relative abundance. Similarly, although there was uneven detection of some taxa at the species and genus levels between replicates, these were only observed at low levels and in a few samples. No other similar studies compared detection but Tap et al. (2019) found no difference on the relative abundance of dominant bacterial taxa, and analysis using DESeq2 found that there were no taxa at the OTU level that were significantly different.

When comparing ecological status metrics, this study found highly significant and strong correlations in TDI values between replicates of the same original sample (Pearson's correlation coefficients ranged from 0.892 to 0.980 with all comparisons  $p < 0.0001$ ) which is similar to the trends observed in specific pollution-sensitivity index (SPI) values by Baricevic et al. (2022), with small, insignificant differences between samples from the same site regardless of the time stored or preservation technique. Unlike Baricevic et al. (2022) who found no differences in Specific Pollution-sensitivity Index classes at the six sites compared, this study found differences in assigned water quality class. Four samples were assigned a greater than one class difference, although the difference was dependent on which replicates were compared. As no species were differentially abundant between replicate groups across all samples,

this suggests that the differences in these outlier replicates are not likely due to fundamental differences caused by sample storage. To elucidate the differences within these outlier samples further replication would be required during the original analysis.

We speculate that the cause of these insignificant but observed differences in sub-sample assemblages and at the taxon levels are due to differences caused by sub-sampling of the original biofilm samples and/or stochastic variation in the communities exacerbated by PCR amplification. Subtle differences caused by sample and PCR variations have been widely reported in the literature and are common caveats of routine monitoring data (Mathieu et al. 2020; Shirazi et al. 2021; Gold et al. 2023). We have come to this conclusion as there was no significant difference in the relative abundance of taxa in samples after storage, and differences in detection of taxa only found in original or repeated analyses were typically in taxa that were detected at low levels and across few samples, and detection of these rare taxa were more likely impacted by sub-sampling. All but two diatom species (*Surirella solea* and *Placoneis constans*) with different levels of detection across replicates belonged to genera that were more widely detected across replicate groups, suggesting that the observed differences in composition were unlikely to be caused by the breakdown of cells or DNA or influenced by the evolutionary relationships (phylogenetic differences) among the species.

Overall, our observed (insignificant) findings suggest that any differences in diatom assemblages are likely masked by variations in the assemblages due to sub-sampling, inter-analyst, and inter-instrument biases, all of which existed between the original and replicate analyses. As a result, diatom assemblages from biofilm samples stored frozen in nucleic acid preservative were not affected by storage for up to three years and are suitable for use in further research. We do, however, suggest caution when extrapolating the results to a wider microbial community because research on the stability of bacterial communities by 16S metabarcoding is contradictory and difficult to compare to the samples used in this study due to differences in sample type and storage conditions (Song et al. 2016; Tap et al. 2019; Delavaux et al. 2020). Little research has been conducted on the impact of storage conditions on fungi or protists (Delavaux et al. 2020). The extent to which DNA from other microbial groups may have degraded in biofilms frozen in preservatives is unknown and should be considered when interpreting the analysis of any big data generated using these archived samples.

This study has further evidenced RNAlater-based preservation of freshwater biofilms, as an alternative to ethanol, one of the standard recommended methods (CEN 2018). Whilst recognising the importance of standardisation to ensure high-quality and comparable data, standardised methods still need to be pragmatic and cost effective in order to ensure uptake by regulatory bodies. For logistic, and health and safety reasons the Environment Agency has restrictions on the use of ethanol, and therefore sought alternatives for the sampling and storage of biofilm samples (Kelly et al. 2018). The results of this study extend the utility of historic raw biofilm samples stored in nucleic acid preservative allowing regulators to maximise their value, by reusing samples for other purposes such as monitoring other non-microbial taxonomic groups (Rivera et al. 2023).

## Conclusion

This study determined the stability of diatom assemblages in biofilm samples analysed before and after storage in a preservative for up to three years using metabarcoding. Differences in diatom assemblages in samples before and after storage were minimal and likely due to bias in sub-sampling of the original samples, as taxa varied equally before and after storage, and (insignificant) differences in beta-diversity were similar to those previously observed when assessing between-analyst and between-instrument variation. This suggests that the diatom assemblages are well preserved within biofilm samples up to 3 years old if stored as pellets in a preservative at -20 °C.

## Additional information

### Conflict of interest

The authors have declared that no competing interests exist.

### Ethical statement

No ethical statement was reported.

### Funding

This work was funded by the Environment Agency under the research project SC220036. The views expressed in this paper are the authors' and do not necessarily represent those of the Environment Agency. DSR was supported by NERC grant NE/X015947/1. JDT was supported by NERC grant NE/X012204/1.

### Author contributions

Jonathan Warren: Conceptualization, methodology, formal analysis, writing - original draft, visualization, funding acquisition; Kerry Walsh: Conceptualization, writing - original draft, funding acquisition; Laura Hunt: Writing - original draft/ review and editing; Sean Butler: Investigation, resources, writing – review and editing; Nick Evens: Resources, writing – review and editing; Joe Taylor: Formal analysis, writing – review and editing; Lindsay Newbold: Formal analysis, writing – review and editing; Dan Read: Writing – review and editing; Martyn Kelly: Writing – review and editing.

### Author ORCIDs

Jonathan Warren  <https://orcid.org/0000-0003-3381-3852>

Sean Butler  <https://orcid.org/0009-0003-4484-5339>

Laura Hunt  <https://orcid.org/0000-0002-4600-5689>

Lindsay Newbold  <https://orcid.org/0000-0001-8895-1406>

Daniel S. Read  <https://orcid.org/0000-0001-8546-5154>

Joe D. Taylor  <https://orcid.org/0000-0003-0095-0869>

Kerry Walsh  <https://orcid.org/0000-0001-8619-8895>

### Data availability

All of the data that support the findings of this study are available in the main text or Supplementary Information.



## References

- Astudillo-García C, Hermans SM, Stevenson B, Buckley HL, Lear G (2019) Microbial assemblages and bioindicators as proxies for ecosystem health status: Potential and limitations. *Applied Microbiology and Biotechnology* 103(16): 6407–6421. <https://doi.org/10.1007/s00253-019-09963-0>
- Baricevic A, Chardon C, Kahlert M, Karjalainen SM, Pfannkuchen DM, Pfannkuchen M, Rimet F, Tankovic MS, Trobajo R, Vasselon V, Zimmermann J, Bouchez A (2022) Recommendations for the preservation of environmental samples in diatom metabarcoding studies, 349–365. <https://doi.org/10.3897/mbmg.6.85844>
- Bisanz JE (2018) qiime2R: Importing QIIME2 artifacts and associated data into R sessions. <https://github.com/jbisanz/qiime2R>
- Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6(1): 1–17. <https://doi.org/10.1186/s40168-018-0470-z>
- Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang K, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciulek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimy AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Priesse E, Rasmussen LB, Rivers A, Robeson MS II, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37(8): 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Bundgaard-Nielsen C, Hagstrøm S, Sørensen S (2018) Interpersonal Variations in Gut Microbiota Profiles Supersedes the Effects of Differing Fecal Storage Conditions. *Scientific Reports* 8(1): 1–9. <https://doi.org/10.1038/s41598-018-35843-0>
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods* 13: 581–583. <https://doi.org/10.1038/nmeth.3869>
- CEN (2018) CEN/TR 17245:2018 Water quality - Technical report for the routine sampling of benthic diatoms from rivers and lakes adapted for metabarcoding analyses. <https://standards.iteh.ai/catalog/standards/cen/31e578ee-1135-49d8-9446-94b56d9c267e/cen-tr-17245-2018>
- Codello A, Hose GC, Chariton A (2022) Microbial co-occurrence networks as a biomonitoring tool for aquatic environments: A review. *Marine and Freshwater Research*: 409–422. <https://doi.org/10.1071/MF22045>
- Cordier T, Lanzén A, Apothéoz-Perret-Gentil L, Stoeck T, Pawlowski J (2019) Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology* 27(5): 387–397. <https://doi.org/10.1016/j.tim.2018.10.012>

- Delavaux CS, Bever JD, Karppinen EM, Bainard LD (2020) Keeping it cool: Soil sample cold pack storage and DNA shipment up to 1 month does not impact metabarcoding results. *Ecology and Evolution* 10(11): 4652–4664. <https://doi.org/10.1002/ece3.6219>
- Deutschmann IM, Lima-Mendez G, Krabberød AK, Raes J, Vallina SM, Faust K, Logares R (2021) Disentangling environmental effects in microbial association networks. *Microbiome* 9: 1–18. <https://doi.org/10.1186/s40168-021-01141-7>
- Dully V, Rech G, Wilding TA, Lanz A, Mackichan K, Berrill I, Stoeck T (2021) Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin* 173: 113129. <https://doi.org/10.1016/j.marpolbul.2021.113129>
- Eastwood N, Zhou J, Derelle R, Abdallah MA-E, Stubbings WA, Jia Y, Crawford SE, Davidson TA, Colbourne JK, Creer S, Bik H, Hollert H, Orsini L (2023) 100 years of anthropogenic impact causes changes in freshwater functional biodiversity. *bioRxiv*: 2023.02.26.530075. <https://doi.org/10.1101/2023.02.26.530075>
- Environment Agency (2023) Using DNA to understand river diatom communities. Environment Agency. [gov.uk](https://www.gov.uk).
- European Parliament (2000) DIRECTIVE 2000/60/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. *Official Journal of the European Union* 43: 1–44.
- European Parliament (2006) Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. *Official Journal of the European Union* 53: 1–12.
- Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive earth's biogeochemical cycles. *Science* 320(5879): 1034–1039. <https://doi.org/10.1126/science.1153213>
- Fontaine L, Pin L, Savio D, Friberg N, Kirschner AKT, Farnleitner AH, Eiler A (2023) Bacterial bioindicators enable biological status classification along the continental Danube river. *Communications Biology* 6(1): 1–11. <https://doi.org/10.1038/s42003-023-05237-8>
- Gold Z, Shelton AO, Casendino HR, Duprey J, Gallego R, Van Cise A, Fisher M, Jensen AJ, D'Agnese E, Allan EA, Ramón-Laca A, Garber-Yonts M, Labare M, Parsons KM, Kelly RP (2023) Signal and noise in metabarcoding data. *PLoS ONE* 18(5): 1–21. <https://doi.org/10.1371/journal.pone.0285674>
- Guseva K, Darcy S, Simon E, Alteio LV, Montesinos-Navarro A, Kaiser C (2022) From diversity to complexity: Microbial networks in soils. *Soil Biology & Biochemistry* 169: 108604. <https://doi.org/10.1016/j.soilbio.2022.108604>
- Jackson MC, Weyl OLF, Altermatt F, Durance I, Friberg N, Dumbrell AJ, Piggott JJ, Tiegs SD, Tockner K, Krug CB, Leadley PW, Woodward G (2016) 55 Advances in Ecological Research Recommendations for the Next Generation of Global Freshwater Biological Monitoring Tools. 1<sup>st</sup> edn. Elsevier, 615–636. <https://doi.org/10.1016/bs.aecr.2016.08.008>
- Kelly M, Juggins S, Guthrie R, Pritchard S, Jamieson J, Rippey B, Hirst H, Yallop M (2008) Assessment of ecological status in U.K. rivers using diatoms. *Freshwater Biology* 53(2): 403–422. <https://doi.org/10.1111/j.1365-2427.2007.01903.x>
- Kelly M, Boonham N, Juggins S, Kille P, Mann DG, Pass D, Sapp M, Sato S, Glover R (2018) Environment Agency A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers, 157 pp.
- Kelly M, Juggins S, Mann DG, Sato S, Glover R, Boonham N, Sapp M, Lewis E, Hany U, Kille P, Jones T, Walsh K (2020) Development of a novel metric for evaluating diatom

- assemblages in rivers using DNA metabarcoding. *Ecological Indicators* 118: 106725. <https://doi.org/10.1016/j.ecolind.2020.106725>
- Kelly MG, Mann DG, Taylor JD, Juggins S, Walsh K, Pitt J-A, Read D (2024) Maximising environmental pressure-response relationship signals from diatom-based metabarcoding in rivers. *The Science of the Total Environment* 914: 169445. <https://doi.org/10.1016/j.scitotenv.2023.169445>
- Kim JH, Jeon JY, Im YJ, Ha N, Kim JK, Moon SJ, Kim MG (2023) Long-term taxonomic and functional stability of the gut microbiome from human fecal samples: 1–8. <https://doi.org/10.1038/s41598-022-27033-w>
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12): 1–21. <https://doi.org/10.1186/s13059-014-0550-8>
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17(1): 10–12. <https://doi.org/10.14806/ej.17.1.200>
- Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL (2020) A systematic review of sources of variability and uncertainty in eDNA data for environmental monitoring. *Frontiers in Ecology and Evolution* 8: 1–14. <https://doi.org/10.3389/fevo.2020.00135>
- McElhinney JMWR, Catacutan MK, Mawart A, Hasan A, Dias J (2022) Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges. *Frontiers in Microbiology* 13: 851450. <https://doi.org/10.3389/fmicb.2022.851450>
- McGuire KL, Treseder KK (2010) Microbial communities and their relevance for ecosystem models: Decomposition as a case study. *Soil Biology & Biochemistry* 42(4): 529–535. <https://doi.org/10.1016/j.soilbio.2009.11.016>
- McMurdie PJ, Holmes S (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8(4): e61217. <https://doi.org/10.1371/journal.pone.0061217>
- McMurdie PJ, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology* 10(4): e1003531. <https://doi.org/10.1371/journal.pcbi.1003531>
- Mishra S, Lin Z, Pang S, Zhang W, Bhatt P, Chen S (2021) Recent Advanced Technologies for the Characterization of Xenobiotic-Degrading Microorganisms and Microbial Communities. *Frontiers in Bioengineering and Biotechnology* 9: 632059. <https://doi.org/10.3389/fbioe.2021.632059>
- Neu TR, Lawrence JR (1997) Development and structure of microbial biofilms in river water studied by confocal laser scanning microscopy. *FEMS Microbiology Ecology* 24(1): 11–25. <https://doi.org/10.1111/j.1574-6941.1997.tb00419.x>
- Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB, Solyomos P, Stevens MHH, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista HBA, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill MO, Lahti L, McGlinn D, Ouellette M-H, Ribeiro Cunha E, Smith T, Stier A, Ter Braak CJF, Weedon J (2022) vegan: Community Ecology Package. <https://cran.r-project.org/package=vegan>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

- R Core Team (2022) R: A Language and Environment for Statistical Computing. <https://www.r-project.org/>
- Rimet F, Gusev E, Kahlert M, Kelly MG, Kulikovskiy M, Maltsev Y, Mann DG, Pfannkuchen M, Trobajo R, Vasselon V, Zimmermann J, Bouchez A (2019) Diat.barcode, an open-access curated barcode library for diatoms. *Scientific Reports* 9(1): 1–12. <https://doi.org/10.1038/s41598-019-51500-6>
- Rivera SF, Vasselon V, Bouchez A, Rimet F (2023) eDNA metabarcoding from aquatic biofilms allows studying spatial and temporal fluctuations of fish communities from Lake Geneva. *Environmental DNA* 5(3): 1–12. <https://doi.org/10.1002/edn3.413>
- Sagova-Mareckova M, Boenigk J, Bouchez A, Cermakova K, Chonova T, Cordier T, Eisen-dle U, Elersek T, Fazi S, Fleituch T, Frühe L, Gajdosova M, Graupner N, Haegerbaeumer A, Kelly AM, Kopecky J, Leese F, Nöges P, Orlic S, Panksep K, Pawlowski J, Petrusek A, Piggott JJ, Rusch JC, Salis R, Schenk J, Simek K, Stovicek A, Strand DA, Vasquez MI, Vrålstad T, Zlatkovic S, Zupancic M, Stoeck T (2021) Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research* 191: 116767. <https://doi.org/10.1016/j.watres.2020.116767>
- Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J (2021) GGally: Extension to “ggplot2.” <https://cran.r-project.org/package=GGally>
- Shirazi S, Meyer RS, Shapiro B (2021) Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. *Ecology and Evolution* 11(22): 15766–15779. <https://doi.org/10.1002/ece3.8239>
- Song SJ, Amir A, Metcalf JL, Amato KR (2016) Microbiome Stability, Affecting. *Msystems.Asm. Org* 1: 1–12. <https://doi.org/10.1128/mSystems.00021-16>
- Tap J, Cools-portier S, Pavan S, Druesne A, Öhman L, Törnblom H, Simren M, Derrien M (2019) Effects of the long-term storage of human fecal microbiota samples collected in RNAlater.: 1–9. <https://doi.org/10.1038/s41598-018-36953-5>
- Warnasuriya SD, Udayanga D, Manamgoda DS, Biles C (2023) Fungi as environmental bioindicators. *The Science of the Total Environment* 892: 164583. <https://doi.org/10.1016/j.scitotenv.2023.164583>
- Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer, New York. [https://doi.org/10.1007/978-3-319-24277-4\\_9](https://doi.org/10.1007/978-3-319-24277-4_9)
- Yan L (2023) ggvenn: Draw Venn Diagram by “ggplot2.” <https://cran.r-project.org/package=ggvenn>
- Yutani H (2022) gghighlight: Highlight Lines and Points in “ggplot2.” <https://cran.r-project.org/package=gghighlight>

## Supplementary material 1

### Extended methods

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl1>



## Supplementary material 2

### PERMANOVA output for 'site' and 'frozen'

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Explanation note: Variable 'frozen' indicated whether the sample was analysed as part of the original analysis or the repeat. Variable 'site' is which original sample and sampling site each replicate is from and is included in the models to account for the expected variation between different locations.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl2>

## Supplementary material 3

### Comparison of differential log2fold abundance between each replicate group

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl3>

## Supplementary material 4

### Distribution of uniquely detected species between sample replicates

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Explanation note: Numbers in parentheses represent number of samples where genera were detected.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl4>

## Supplementary material 5

### Taxonomic bar plots of diatom taxa at the order level of the 4 outlier TDI samples 09, 43, 51, and 55

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl5>

## Supplementary material 6

### Taxonomic bar plots of diatom taxa at the order level of all samples

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: pdf

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl6>

## Supplementary material 7

### Merged data and metadata

Authors: Jonathan Warren, Sean Butler, Nick Evens, Laura Hunt, Martyn Kelly, Lindsay Newbold, Daniel S. Read, Joe D. Taylor, Kerry Walsh

Data type: xlsx

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.8.129227.suppl7>